Available online at www.sciencedirect.com

Atmospheric Environment 40 (2006) 5759–5767

ATMOSPHERIC ENVIRONMENT

www.elsevier.com/locate/atmosenv

Off-line diagnostic analyses of a three-dimensional PM model using two matrix factorization methods

Jinyou Liang[a,*], Ajith Kaduwela[a,b], Bruce Jackson[a], Kemal Gürer[a], Paul Allen[a]

[a]Air Resources Board, California Environmental Protection Agency, 1001 I Street, Sacramento, CA 95814, USA
[b]Department of Land, Air, and Water Resources, University of California at Davis, Davis, CA, USA

Received 26 October 2005; received in revised form 28 April 2006; accepted 11 May 2006

Abstract

Error diagnosis of fine-grid photochemical transport models (CTM) has become a formidable task, which requires thorough understanding of complex microphysical and photochemical processes in the atmosphere as well as scientific computing. In an initial modeling exercise conducted for the California Regional $PM_{10}/PM_{2.5}$ Air Quality Study (CRPAQS), abnormally high, unrealistic, PM sulfate concentrations were simulated in central California. To aid the error diagnosis, two matrix factorization methods, namely absolute principal component analysis (APCA) and an efficient non-negative matrix factorization method (NMFROC), were used to analyze the relationships among the input and output parameters of a CTM for PM modeling and to apportion the relative importance of individual factors to an abnormal sample. The APCA method corroborated sciences implemented in the PM model, but failed to apportion the relative importance of individual factors to PM sulfate in an abnormal case. On the other hand, the NMFROC method performed well on the apportionment of an abnormally high PM sulfate. The factors produced from the NMFROC method shared common features with the APCA method, but significant differences remain between the two methods, which can be understood from their difference in methodology. Subsequent PM modeling results were shown to validate the results from the NMFROC method.
Published by Elsevier Ltd.

Keywords: Corroborative analysis tool; Air quality model; Sulfate; San Joaquin Valley

1. Introduction

Grid-based photochemical transport models, such as CMAQ (USEPA, 1999; CMAS, 2005) and CAMx (Environ, 2005), require inputs of three-dimensional meteorological parameters and emission rates of gas and particle species, to generate outputs for concentrations of chemical species and particle parameters. Owing to the large volume of input and output data, the error diagnosis of PM models has become a formidable task. Implementation of process analysis in PM models was shown to provide helpful information on certain processes (Tonnesen and Dennis, 2000), and sensitivity analysis tools may be also helpful (Morris et al., 2003; Zhang et al., 2005b). However, these techniques have to be run on-line, which further slows down 3D PM models that are already computationally demanding (Zhang et al., 2005a). A complementary off-line diagnostic tool is desirable

*Corresponding author. Tel.: +1 916 322 1223; fax: +1 916 327 8524.
E-mail address: jliang@arb.ca.gov (J. Liang).

1352-2310/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.atmosenv.2006.05.035

especially if the computer resource is an issue as for the 2000–2001 wintertime PM modeling in the California Regional $PM_{10}/PM_{2.5}$ Air Quality Study (Liang et al., 2006a; Magliano and McDade, 2005; Zhang et al., 2005a).

Receptor-oriented models have been previously applied to addressing source identification and apportionment issues of water and air pollution (Winchester and Nifong, 1971; Henry et al., 1984; Hopke, 1985; Watson et al., 1990, 2001; Chow and Watson, 2002; Lewis et al., 2003). For receptor models, the application problems have to be linear, and no significant change is allowed for source profiles between the emission and receptor points. Inert or slow-reacting primary pollutants, such as elements and CO, are about linear in terms of source apportionment between sources and receptors. Photochemical products, such as ozone and secondary PM, are non-linear in terms of source apportionment, since their responses at receptors to precursor reductions at sources are often not proportional. Meteorological parameters are also non-linear in nature, since they are non-additive and source apportionment is irrelevant for them. In sum, for non-linear species and parameters, the source apportionment function of receptor models is meaningless. Matrix factorization (MF) methods used in receptor-oriented models, however, could be used to analyze the relationship between model inputs and outputs, to be elaborated below, since the mathematical algorithms of MF methods, such as principal component analysis (PCA) (Thurston and Spengler, 1985; Jolliffe, 2002) and non-negative MF (NMF) methods (Paatero, 1997; Lee and Seung, 1999, 2001; Liang and Fairley, 2006), were designed for broader applications. Using NMF as an example, its mathematical goal is to extract a number of extreme rays (or called parts, components, vectors, etc.) in the positive orthant from sample matrix to account for major features of the sample matrix. The NMF method carries no assumption to or inference from the information before the data were acquired. Thus, it leaves the interpretation of results to users in specific fields according to the properties of the sample matrix and the nature of the NMF method.

To simulate an extended 2000–2001 winter PM episode captured in the Central Valley during the California Regional $PM_{10}/PM_{2.5}$ Air Quality Study (CRPAQS), we conducted a series of simulations using CMAQ with MM5 meteorological inputs. Details about the CMAQ simulations for the above CRPAQS episode (Liang et al., 2006a,b; Zhang et al., 2005a) are not the focus of this paper. In earlier simulations, abnormally high, unrealistic concentrations of PM sulfate were produced in the model. We applied two MF methods to aid in error diagnosis, as well as corroborate model performance. An efficient non-negative matrix factorization method (NMFROC) (Liang and Fairley, 2006) and the absolute PCA method (Thurston and Spengler, 1985; Cao et al., 2005) were coded in a statistical language (R Development Core Team, 2005). First, we will briefly introduce the two MF methods in Section 2. Then, we will describe the PM modeling problem and parameters in Section 3. After that, we will present the results from MF methods in Section 4. Finally, we will conclude with a summary.

2. The two matrix factorization methods

In this section, we will briefly describe the two MF methods used in this paper. For more detailed formulation, readers are referred to Thurston and Spengler (1985) for absolute PCA (APCA) and Liang and Fairley (2006) for NMFROC.

2.1. The APCA method

PCA has been widely used in many fields (Jolliffe, 2002). PCA makes use of eigenvectors of the correlation matrix of input data matrix A with v variables and s samples, to split normalized input matrix $\mathring{A}$ (2.1) into two matrices, namely, an eigenvector matrix $D[v,v]$ that is also termed PC coefficients, and a PC score matrix $(D^t\mathring{A})$. It is common practice to discard those eigenvectors with eigenvalues less than 1, so that only p $(<v)$ factors are retained. The APCA method rotates the $D[v,p]$ matrix with a scheme called varimax to reach a final coefficient matrix D^*, and calibrates the corresponding PC score matrix $(S=D^{*t}\mathring{A})$ to reach the absolute PC score matrix X, as shown in Eq. (2.2). For factor identification purposes, the correlation between variables and PCs in the samples was calculated to form a PC loading matrix. X can be used in subsequent regression against variables of interest related to samples.

$$\mathring{A}[iv,is]=\frac{A[iv,is]-\bar{A}[iv]}{\sigma[iv]},$$
$$is=1:s,\quad iv=1:v, \tag{2.1}$$

$$X[ip, is] = S[ip, is] + \sum_{iv=1}^{v} D^{*}[iv, ip]\frac{\bar{A}[iv]}{\sigma[iv]},$$
$$is = 1:s,\ \ ip = 1:p. \tag{2.2}$$

2.2. The NMFROC method

The NMFROC method is relatively new, and the detailed formulation was presented in Liang and Fairley (2006). The NMFROC method consists of two major steps, namely the initialization step (ROC) and the updating step (NMF). For an input matrix A with v variables and s samples, the ROC method obtains initial solution matrices ($B[v,p]$ and $C[p,s]$) from $A[v,s]$, where $p<v<s$, with non-negative constraints. The updating step proceeds until the solution converges to either absolute error tolerance ($\mathrm{PE}\leqslant 1^{\circ}$) or relative error tolerance ($\mathrm{PE}_n/\mathrm{PE}_{n-1}\leqslant 10^{-7}$), where PE was defined by Liang and Fairley (2006). After convergence, columns of B are scaled so that the sum of column elements equals v. The rows of C are scaled accordingly, so that the product of B and C and source apportionment results stay unchanged.

3. The PM modeling problem and the formation of the input matrix

In this section, we first describe the problem that we encountered in CRPAQS PM modeling, and then illustrate the procedures taken to form input matrix for the NMFROC method.

3.1. The PM modeling problem

PM episodes were frequent in the winter until recent years in the Central Valley, which covers metropolitan Fresno and Bakersfield and surrounding rural areas (Fig. 1). A 2-week PM episode was captured in the Central Valley during 25 December 2000 throughout 7 January 2001 (Magliano and McDade, 2005). To simulate this episode, we configured the CMAQ with the horizontal domain that covers central and northern California and adjacent areas, shown in Fig. 1. The model contains (185 × 185) (4 km × 4 km) horizontal grids and 15 vertically expanding, terrain-following sigma-P layers, with the bottom layer thickness of ~30 m and the top layer reaching 10 000 Pa. The meteorological inputs were generated from NCAR/PSU Mesoscale Meteorological Model (version 5), with three one-way nested grids. In the two parent grids, analysis-nudging was applied, but no FDDA was applied in the third grid used in this study. Emission inputs were speciated and gridded from regulatory emission inventory maintained at California Air Resources Board (CARB) for a larger modeling effort, the CRPAQS PM modeling study, conducted by scientists from multi-agencies. Several dozens of simulations were conducted at CARB to find out the scientific causes for the PM episode (Liang et al., 2006a,b). In earlier simulations, abnormally high, unrealistic concentrations of fine PM sulfate were produced in the Central Valley, as shown in Fig. 2(a).

3.2. Formation of the input matrix

To diagnose the error(s) in the model parameters and/or processes responsible for the above problem and to corroborate model sciences, we employed the APCA and NMFROC methods described in the last section. The procedures for conducting APCA analysis can be found in a number of literatures (Thurston and Spengler, 1985; Cao et al., 2005). We focus here on the steps taken to conduct NMFROC analysis for grid-based PM models.

For both MF methods, the input data matrix consists of hourly quantities of emissions, simulated concentrations of trace gases and fine PM components, and important meteorological parameters at three anchor stations of the CRPAQS during 25 December 2000 throughout 7 January 2001. The deposition parameters used in the PM model were excluded from the input matrix for the NMFROC method, since they were not expected to have significant impacts on the PM sulfate anomaly. The emissions in adjacent grids may play important roles in concentrations at the anchor sites, but we decided to use the emissions in corresponding grids as surrogates for all emissions to keep the input matrix at manageable size for this study. The resulting input matrix contained 76 variables (Table 1) and 1008 (24 × 14 × 3) hourly samples, which led to the total number of elements to be 76608.

For the APCA method, the input matrix was normalized to have the standard (0,1) distribution for all variables. For the NMFROC method, several steps were taken to convert a raw species from CMAQ inputs and outputs to a variable in the input matrix (A). First, each species was subtracted with its minimum value. This step enabled the inclusion of meteorological variables, such as wind

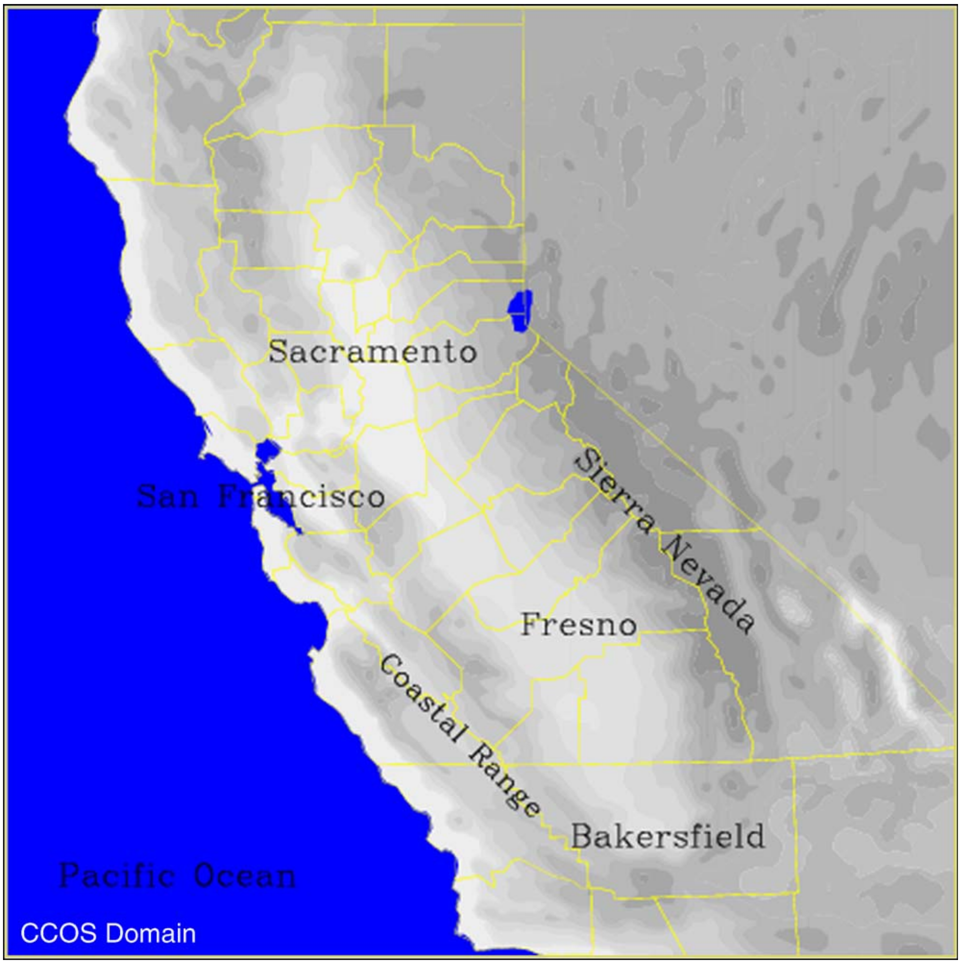

Fig. 1. Horizontal model domain. Fresno and Bakersfield are the two major cities in the Central Valley.

components and the inverse of the Monin–Obukov length (MOLI) that can be negative. Second, each species was then divided by its mean quantity. This step removed the units of species, thus allows for the association among transformed variables. We understand that standard deviation can also be used to substitute the mean in the second step, but it will not change apportionment results. The above two steps preserved the monotonic relationship between the original variables and the corresponding variables in the input matrix, a desirable feature for interpretation of the NMFROC results. We understand that the ranges, not the signs, of winds and MOLI were preserved in these steps. If necessary, the absolute quantities of these variables can be recovered in the similar manner as APCA. Since the minimum PM sulfate was much smaller than the maximum in this study, it can be proved that the above steps have negligible impact on apportionment results for PM sulfate.

4. Results and discussions

4.1. The APCA results

We conducted PCA analysis on the input data matrix, using the standard normalization and singular value decomposition techniques written in a statistical language (R Development Core Team, 2005). Nine principal components (PC) with eigenvalues larger than 1 were retained, and subjected to varimax rotation. The coefficients of the nine PCs, together with their eigenvalues and correlation coefficients with variables, or called loadings, in the input matrix, are listed in Table 2. We briefly summarize the features of the nine PCs here.

Factors 1–2 indicate clustering of emissions and primary species. Factor 3 reflects that high ozone was related to rural air with high PAN, NH_3, dust, and low primary anthropogenic PM emissions. Factor 4 reflects secondary production of organic PM and nitrate was associated with high temperature and H_2O_2. Factor 5 reflects that PM sulfate was associated with PM water content, as well as ammonium, cloud fraction, low temperature and H_2O_2. Factor 6 reflects that ozone and HNO_3 were high in unstable conditions (low MOLI) when PBL and ground radiation were high. Factor 7 associates ALK1 and crustal PM in the air with emissions. Factor 8 associates cloud fraction and rain-water with northwestern winds. Factor 9 reflects that high H_2O_2 was associated with northern, downward flows. The above factors are consistent with our understanding of the science related to ozone and PM in the model. If PC loading alone is considered, then Factor 5 suggests that PM sulfate was highly associated with PM water only. Note that PCA tends to extract average information from samples, and the varimax rotation tends to push the results towards extreme rays, the direction of the NMFROC method (Thurston et al., 2005; Liang and Fairley, 2006).

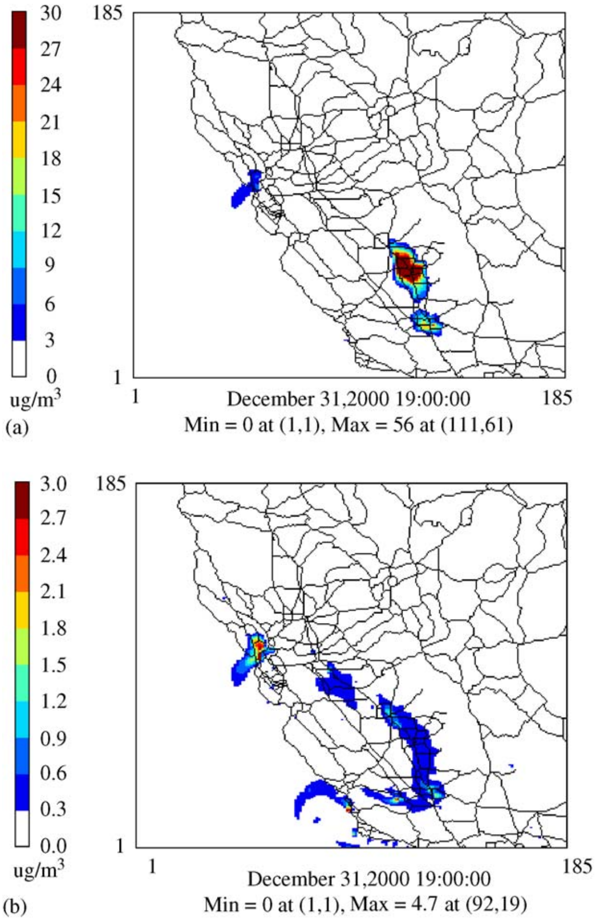
30
27
24
21
18
15
12
9
6
3
0
ug/m³
185
1
1
185
(a)
December 31,2000 19:00:00
Min = 0 at (1,1), Max = 56 at (111,61)
3.0
2.7
2.4
2.1
1.8
1.5
1.2
0.9
0.6
0.3
0.0
ug/m³
185
1
1
185
(b)
December 31,2000 19:00:00
Min = 0 at (1,1), Max = 4.7 at (92,19)

Fig. 2. Sulfate in the fine PM mode simulated with MM_5 rains/clouds (a, upper panel), and without rains but with PM liquid water content within $0.1\,g\,m^{-3}$ (b, lower panel), at 11 am, 31 December 2000.

We calculated the absolute PC scores following Thurston and Spengler (1985), and fitted hourly PM sulfate with the scores at the three anchor sites during the winter PM episode with and without the intercept, using an ordinary regression scheme (lm) implemented in *R*. The regressed formula, in combination with the PC scores, was used to calculate contributions from nine factors to the observed PM sulfate in Bakersfield at the time shown in Fig. 2a. The results were listed in the last column of Table 5. It is shown that calculations using the APCA method produced large, negative values for contributions from individual components, though the larger to smaller (L/S) ratio (Liang and Fairley, 2006) was 1.29. The exclusion of the intercept in the regression analysis using the APCA method, not shown here, resulted in larger L/S ratio but did not eliminate large, negative terms. The apportionment results could be improved by

Table 1
Variables used in input matrix

Concentration: gas	O_3, NO, NO_2, PAN, CO, HNO_3, SO_2, HCHO, HO_2H, ALK1, ALK2, ALK3, ALK4, ALK5, ARO1, ARO2, OLE1, OLE2, ETHENE, TRP1, NH_3
Concentration: $PM_{2.5}$	AH_2O, ANH_4, ANO_3, ASO_4, AEC, AORGPA, A25, AORGA, AORGB, NUMACC, SRFACC
Emission: gas	NO, NO_2, HONO, SO_2, SULF, HCHO, CCHO, RCHO, ACET, MEK, CH_4, ETHENE, ISOPRENE, TRP1, MTBE, ETOH, ALK1, ALK2, ALK3, ALK4, ALK5, ARO1, ARO2, OLE1, OLE2, CO, NH_3
Emission: $PM_{2.5}$	PMFINE, PEC, POA, PNO_3, PSO_4, Dp, σ
Meteorological parameters	PBL, MOLI, RGRND, CFRAC, WWIND, TA, QR, QC, UWIND, VWIND

Note: Variable names followed CMAQ nomenclatures, except for D_p and σ that were added for CRPAQS study. D_p is the geometric mean volume diameter, and σ the geometric standard deviation.

Table 2
Key coefficients and loadings from APCA with varimax rotation

Factor	Eigenvalue	Variables with large PC coefficients and high loadings
1	35.0	Emis: NO (0.22, **0.95**), NO_2(0.22, **0.95**), HONO (0.22, **0.95**), SO2(**0.72**), SULF(**0.72**), HCHO(0.21, **0.98**), CCHO(0.18, **0.97**), RCHO(**0.85**), ACET(**0.76**), MEK(**0.91**), C_2H_4(0.24, **0.99**), ISOP(**0.86**), TRP1(**0.78**), MTBE(0.22, **0.95**), ALK2(0.20, **0.97**), ALK3(**0.92**), ALK4(0.20, **0.96**), ALK5(**0.92**), ARO1(0.21, **0.98**), ARO2(0.22, **0.98**), OLE1(0.23, **0.98**), OLE2(0.24, **0.96**), CO(0.22, **0.96**), PEC(0.21, **0.94**), POA(**0.79**)
2	13.1	Chem: NO(0.28, **0.87**), NO_2(0.20, **0.89**), CO(0.24, **0.97**), SO_2(0.22, **0.89**), HCHO(0.27, **0.89**), ALK1(**0.74**), ALK2(0.24, **0.97**), ALK3(0.22, **0.96**), ALK4(0.23, **0.98**), ALK5(0.21, **0.95**), ARO1(0.25, **0.98**), ARO2(0.24, **0.97**), OLE1(0.24, **0.97**), OLE2(0.26, **0.96**), C_2H_4(0.24, **0.97**), TRP1(**0.88**), EC(0.25, **0.98**), AORGPA(0.18, **0.91**), A25(**0.81**), NUMACC(**0.76**), SRFACC(**0.86**)
3	5.42	Chem: O_3 (0.20), PAN(0.19), NH_3(0.28) Emis: NH_3(0.42, **0.90**), PNO_3(−0.36, −**0.95**), PSO_4(−0.27, −**0.88**), Dm(0.31, **0.87**), σ(0.28, **0.71**)
4	3.38	Chem: H_2O_2 (0.26), ANH_4(0.29), ANO_3(0.44, **0.74**), AORGA(0.42, **0.86**), AORGB(0.43, **0.74**) Met.: TA(0.29)
5	3.23	Chem: H_2O_2(−0.21), ANH_4(0.44, −**0.75**), AH_2O(0.45, **0.85**), ASO_4(0.52, **0.88**) Met: QC(0.31), TA(−0.24)
6	2.31	Chem: O_3(0.31), HNO_3(0.23), NH_3(−0.24) Emis: σ(0.24) Met.: PBL(0.41, **0.79**), MOLI(−0.39, −**0.82**), RGRND(0.46, **0.89**)
7	1.85	Chem: CO(**0.76**), ALK1(0.25, **0.76**), ALK4(**0.75**), ALK5(**0.75**), ARO1(**0.73**), ARO2(**0.76**), OLE1(**0.74**), OLE2(**0.71**), C_2H_4(**0.73**), A25(0.17), NUMACC(0.19, **0.87**), SRFACC(**0.75**) Emis: SO_2(**0.83**), SULF(**0.83**), CCHO(**0.76**), RCHO(0.20, **0.93**), ACET(0.25, **0.95**), CH_4(0.29, **0.96**), TRP1(0.23, **0.92**), MTBE(**0.76**), ETOH(0.28, **0.94**), ALK1(0.30, **0.93**), ALK2(**0.70**), ALK3(**0.76**), ALK4(**0.80**), ARO2(**0.70**), OLE1(**0.72**), OLE2(**0.72**), CO(**0.76**), PMFINE(0.25, **0.94**), POA(0.21, **0.92**)
8	1.43	Chem: AH_2O(−0.20) Met.: CFRAC(−0.62, −**0.83**), QR(−0.26), QC(−0.33, **0.61**), UWIND(0.40), VWIND(−0.29)
9	1.32	Chem: H_2O_2(0.22) Met.: WWIND(−0.74, −**0.87**), VWIND(−0.57, −**0.72**)

Note: $|\text{Coef.}|\geqslant 1.5\sqrt{v}$ are shown, and $v = 76$ in this study. 88% of variance was explained by the first nine factors listed here. Correlation (R) between variables and PCs is listed in bold numbers when $R\geqslant 0.7$.

applying a non-negative fitting to the APCA scores, similar to the ROC method (Liang and Fairley, 2006), but further work is beyond the scope of this paper.

4.2. The NMFROC results

We ran the NMFROC model to obtain nine extreme rays that roughly enclose the ensemble of samples. Runs with fewer extreme rays were also conducted. However, runs with more extreme rays were not conducted due to the constraints of the ROC method and considerations of the prediction error that are already low. Below we illustrate the steps taken to interpret the results.

Table 3 lists the variables with coefficients larger than 2 in the nine extreme rays (columns of B), together with factor loadings $\geqslant 0.7$. Variables with values between 1 and 2 in B were listed in brackets only when the number of variables with larger values was too few. The larger the value of a variable is in an extreme ray, the closer the distance between the corresponding variable and the extreme ray (factor). It is shown that some factors correspond to unique processes, such as rain (factor 1), fog (factor 3), anthropogenic (diesel) emissions (factor 2), and chemical concentrations (factor 4). Other factors combine several processes. For example, factor 7 features MTBE and OLE2 as well as other anthropogenic and isoprene emissions. Factor 5 represents conditions with relatively strong solar insolation, deep PBL height, large biogenic emissions, and large concentrations of nitric acid and ozone as well as hydrogen peroxide. Factor 6 associates high concentration of sulfate in fine PM with very high aerosol water content, and with terpene and organic PM to a much lesser degree. Factor 8 indicates that high cloud fraction is associated with elevated PBL and surface concentrations of ozone and PAN. Factor 9 associates high ozone with a stable, rural condition with large emissions of ammonia and coarser portion of fine PM as well as a number of other variables.

Table 3
Key coefficients and loadings from the NMFROC method

Factor	Variables with high values in matrix *B*
1	QR(65.4, **1.0**) [CFRAC(1.6), PBL(1.5)]
2	Emis: NO(**0.81**), NO2(**0.81**), HONO(**0.81**), NH3(−**0.89**), PNO$_3$(3.8, **0.93**),PSO$_4$(2.8, **0.73**), PEC(2.4), ALK2(2.3), ALK3(2.0), ALK5 (2.2), HCHO(2.8), NO$_x$(3.4)
3	QC(42.7, **1.0**), CFRAC(7.7), AH$_2$O(12.3, **0.73**)
4	Chem: VOC(2.3−3.7), NO(3.3, **0.83**), NO$_2$(2.0, **0.83**), CO(2.8, **0.91**), SO$_2$(3.4, **0.90**), HCHO(2.3, **0.80**), ALK1(**0.78**), ALK2(**0.94**), ALK3(**0.94**), ALK4(**0.95**), ALK5(**0.87**), ARO1(**0.91**), ARO2(**0.90**), OLE1(**0.92**), OLE2(**0.91**), C$_2$H$_4$(**0.91**), AEC(2.4, **0.92**), AORGPA(3.2, **0.94**), A25(2.1, **0.86**), SRFACC(**0.80**)
5	Met: RGRND(12.2, **0.83**), PBL(9.9, **0.72**) Emis: ISOPRENE(5.7, **0.63**), ALK5(2.0), (σ(1.7), PEC(1.7)) Chem: HNO$_3$(16.2, **0.88**), O$_3$(4.7), [HO$_2$H(1.7)]
6	Chem: SRFACC(2.7), AORGB(2.3), AORGPA(2.0), AH$_2$O(21.8, **0.92**), ASO$_4$(15.3, **0.85**), TRP1(2.0)
7	Emis: SO$_2$(2.1, **0.85**), SULF(2.1, **0.85**), CCHO(**0.89**), RCHO(**0.93**), ACET(2.0, **0.92**), MEK(**0.79**), CH$_4$(2.0, **0.87**), ETHENE(2.2, **0.88**), ISOPRENE(2.0, **0.74**), TRP1(**0.81**), MTBE(2.4, **0.91**), ETOH(2.2, **0.92**), ALK1(2.3, **0.87**), ALK2(**0.75**), ALK3(**0.75**), ALK4(2.1, **0.90**), ARO1(2.0, **0.82**), ARO2(2.1, **0.86**), OLE1(2.3, **0.90**), OLE2(2.4, **0.91**), CO(2.3, **0.91**), PMFINE(2.1, **0.94**), POA(**0.81**)
8	Chem: O$_3$(3.9), PAN(2.4); PBL(5.0), CFRAC(41.5, **0.93**)
9	Chem: O$_3$(6.5), PAN(4.1), HO$_2$H(3.2), NH$_3$(5.8), ANH$_4$(3.1), ANO$_3$(3.9), AORGA(2.4), AORGB(2.8), NUMACC(−**0.81**) Emis: NO(−**0.77**), NO$_2$(−**0.77**), HONO(−**0.77**), HCHO(−**0.80**), CCHO(−**0.76**), RCHO(−**0.74**), MEK(−**0.71**), TRP1(−**0.86**), ALK2(−**0.83**), ALK3(−**0.87**), ALK4(−**0.74**), ALK5(−**0.71**),ARO1(−**0.72**), NH$_3$(5.3), POA(−**0.86**), PNO$_3$(−**0.74**), PSO$_4$(−**0.84**), Dp(7.9, **0.92**), σ(4.5) Met: MOLI(3.62), WWIND(3.2), TA(2.9), UWIND(3.1), VWIND(3.1)

Note: Variable names follow Table 1. Variables were listed when their values in *B* (in parenthesis) are larger than 2, while 1 is the average value. A few variables with values smaller than but close to 2 were listed in brackets. Correlation (*R*) between variables and factors is listed in bold numbers when $R \geqslant 0.7$.

Table 4
Correlations between factors resolved from the NMFROC and APCA methods

NMF no.	1	2	3	4	5	6	7	8	9
PC no.	8	3	5, 8	2	6	5	1, 7	8	3, 7
R	0.39	−0.85	−0.64, 0.61	0.95	0.84	−0.90	0.82,−0.86	0.68	0.85, 0.73

Table 4 lists the correlations between factors identified by the NMFROC and APCA methods. It is shown that these two methods identified some common factors, but significant differences existed in other factors, presumably owing to the difference in methodology (Liang and Fairley, 2006).

Fig. 3 shows the PM sulfate with concentrations larger than 1 $\mu g\,m^{-3}$ in the input (black circles) and prediction matrices (blue lines), and the factors that contributed more than 20% to the PM sulfate with concentrations larger than 4 $\mu g\,m^{-3}$, a typical observed value during the PM episode captured in San Joaquin Valley (Liang et al., 2006b). It is shown that factor 6 was dominant in all the above cases, with minor contributions from factor 7. Thus, the PM liquid water content appeared to be the major factor for the PM sulfate anomaly.

For the fine PM sulfate shown in Fig. 2(a), Table 5 lists the corresponding column of *C* and row of *B*, together with products of the two, scaled to 100. The third-to-last column of Table 5 mimics the percentage contributions from factors to sulfate in this sample, with the prediction to observation ratio of 1.28. We refrain here from drawing too much attention to accurate accounting of individual variables. Instead, we intend to qualitatively define the processes that are responsible for the peak sulfate. From the discussion on the features of factors above, peak sulfate in Fig. 2(a) was mainly due to a factor with high aerosol water content, with small influences from factors featuring biogenic and anthropogenic emissions.

To verify the information offered by the NMFROC analysis, we show in Fig. 2(b) the fine

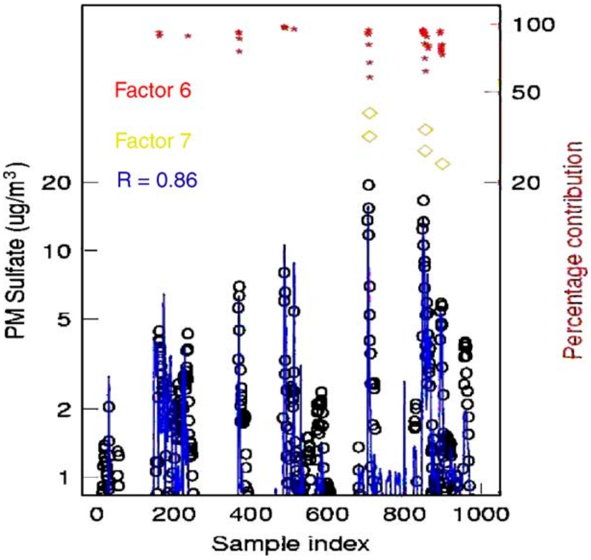

Fig. 3. PM sulfate in the input and prediction matrices and major contributing factors in samples with PM sulfate larger than 4 μg m^{-3}.

Table 5
Apportionment of PM sulfate in a sample from NMFROC and APCA

NMFROC				APCA	
Factor	*C* ["peak"]	*B* ["ASO4"]	*B* × *C* (%)	Factor	Intercept 402
1	0.0	0.1	0.0	PC1	16
2	0.5	0.4	6.4	PC2	7.2
3	0.0	0.3	0.1	PC3	0.1
4	0.2	0.2	1.1	PC4	−30
5	0.1	0.0	0.0	PC5	−309
6	0.2	15.3	87.5	PC6	13
7	0.2	0.7	5.1	PC7	−3.6
8	0.1	0.0	0.0	PC8	−5.2
9	0.0	0.7	0.0	PC9	9.7

PM sulfate in a simulation without rains and with aerosol water content limited to be within 0.1 g m^{-3}. In addition, the catalytic pathways that produces sulfate in aqueous phase but had no dependence on other chemical reactions were turned off. The resulting sulfate in the fine PM mode was reduced by ~10-fold, consistent with the results shown in Fig. 3 and Table 5. Thus, the NMFROC diagnostic results on the abnormal sulfate in fine PM mode were verified.

5. Summary

Error diagnosis of fine-grid photochemical transport models (CTM) has become a formidable task, which requires thorough understanding of complex microphysical and photochemical processes in the atmosphere as well as scientific computing. In an initial modeling exercise conducted for the California Regional $PM_{10}/PM_{2.5}$ Air Quality Study, abnormally high, unrealistic, PM sulfate concentrations were simulated in central California. To aid the error diagnosis, two matrix factorization methods, namely APCA and an efficient NMFROC, were used to analyze the relationships among the input and output parameters of a CTM for PM modeling and to apportion the relative importance of individual factors to PM sulfate in an abnormal sample. The APCA method corroborated sciences implemented in the PM model, but failed to apportion the relative importance of individual factors to PM sulfate in an abnormal sample. On the other hand, the NMFROC method performed well on the apportionment of an abnormally high PM sulfate, as well as identifying the problem in the PM model. The factors produced from the NMFROC method shared some common features of, but showed significant difference from the APCA method, which can be understood from their difference in methodology. Subsequent PM modeling results were shown to validate the results from the NMFROC method.

Acknowledgements

We thank an anonymous reviewer for providing insightful comments that significantly improved the quality of this paper. We are indebted to Professor Michael Kleeman at University of California at Davis for speciating PM emissions in discrete size bins, Professor Thurston at New York University for his guidance on the APCA, and Mr. Donald Johnson at CARB for proofreading part of the text.

Disclaimer: This paper has been reviewed by the staff of the California Air Resources Board and has been approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the California Air Resources Board, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

References

Cao, J.J., Chow, J.C., Lee, S.C., Li, Y., Chen, S.W., An, Z.S., Fung, K., Watson, J.G., Zhu, C.S., Liu, S.X., 2005. Characterization and source apportionment of atmospheric organic and elemental carbon during fall and winter of 2003 in Xi'an, China. Atmospheric Chemistry and Physics Discussion 5, 3561–3593.

Chow, J.C., Watson, J.G., 2002. Review of $PM_{2.5}$ and PM_{10} apportionment for fossil fuel combustion and other sources by the Chemical Mass Balance receptor model. Energy & Fuels 16 (2), 222–260.

Community Modeling and Analysis System (CMAS), 2005. Community Multiscale Air Quality Modeling System (CMAQ), http://www.cmascenter.org.

ENVIRON, 2005. Comprehensive Air Quality Model with Extensions (CAMx); http://www.camx.com.

Henry, R.C., Lewis, C.W., Hopke, P.K., Williamson, H.J., 1984. Review of receptor model fundamentals. Atmospheric Environment 18, 1507–1515.

Hopke, P.K., 1985. Receptor Modeling in Environmental Chemistry. Wiley, New York.

Jolliffe, I.T., 2002. Principal Component Analysis, second ed. Springer, New York.

Lee, D.D., Seung, H.S., 1999. Learning the parts of objects with nonnegative matrix factorization. Nature 401, 788–791.

Lee, D.D., Seung, H.S., 2001. Algorithms for non-negative matrix factorization. Advances in Neural Information Processing Systems 13, 556–562.

Lewis, C.W., Norris, G.A., Henry, R.C., Conner, T.L., 2003. Source Apportionment of Phoenix PM-2.5 Aerosol with the Unmix Receptor Model. Journal of the Air & Waste Management Association 53, 325–338.

Liang, J., Fairley, D., 2006. Validation of an efficient non-negative matrix factorization method and its preliminary application in central California. Atmospheric Environment 40, 1991–2001.

Liang, J., Kaduwela, A., Zhang, K.M., Gurer, K., Ying, Q., Allen, P.D., Kleeman, M., Wexler, A.S., Magliano, K., Jackson, B., DaMassa, J., 2006a. A photochemical model investigation of an extended winter PM episode captured in central California: I. Sensitivity simulations. Atmospheric Environment (to be submitted).

Liang, J., Gurer, K., Kaduwela, A., Zhang, K.M., Ying, Q., Allen, P.D., Kleeman, M., Wexler, A.S., Magliano, K., O'Brien, G., Turkiewicz, K., DaMassa, J., 2006b. A photochemical model investigation of an extended winter PM episode captured in central California: II. Model performance evaluation. Atmospheric Environment (to be submitted).

Magliano, K., McDade, C.E., 2005. The California Regional PM_{10}/$PM_{2.5}$ Air Quality Study (CRPAQS): field study description and initial findings. California Air Resources Board, Sacramento, CA.

Morris, R.E., Yarwood, G., Emery, C.A., Koo, B., 2003. Development and application of the CAMx regional one-atmospheric model to treat ozone, particulate matter, visibility, air toxics and mercury. In: The 96th Annual Conference of Air & Waste Management Association, San Diego, CA, USA, 22–26 June 2003.

Paatero, P., 1997. Least square formulation of robust non-negative factor analysis. Chemometrics Intellectual Laboratory Systems 37, 23–35.

R Development Core Team, 2005. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; URL http://www.R-project.org. ISBN:3-900051-00-3.

Thurston, G.D., Spengler, J.D., 1985. A quantitative assessment of source contributions to inhalable particulate matter pollution in metropolitan Boston. Atmospheric Environment 19, 9–25.

Thurston, G.D., Ito, K., Mar, T., Christensen, W.F., Eatough, D.J., Henry, R.C., Kim, E., Laden, F., Lall, R., Larson, T.V., Liu, H., Neas, L., Pinto, J., Stolzel, M., Suh, H., Hopke, P.K., 2005. Workgroup report: workshop on source apportionment of particulate matter health effects—intercomparison of results and implications. Environmental Health Perspectives 113 (12), 1768–1774.

Tonnesen, G.S., Dennis, R.L., 2000. Analysis of radical propagation efficiency to assess ozone sensitivity to hydrocarbons and NO_x, Part 1: local indicators of odd oxygen production sensitivity. Journal of Geophysical Research 105, 9213–9225.

USEPA, 1999. Science Algorithms of the EPA Models-3 Community Multi-scale Air Quality (CMAQ) Modeling System. In: Byun, D.W., Ching, J.K.S. (Ed.), EPA Report, EPA/600/R-99/030, NERL, Research Triangle Park, NC.

Watson, J.G., Robinson, N.F., Chow, J.C., Henry, R.C., Kim, B.M., Pace, T.G., Meyer, E.L., Nguyen, Q., 1990. The USEPA/DRI Chemical Mass Balance receptor model, CMB7.0. Environmental Software 5, 38–49.

Watson, J.G., Chow, J.C., Fujita, E.M., 2001. Review of volatile organic compound source apportionment by chemical mass balance. Atmospheric Environment 35, 1567–1584.

Winchester, J.W., Nifong, G.D., 1971. Water pollution in Lake Michigan by trace elements from pollution aerosol fallout. Water, Air and Soil Pollution 1, 50–64.

Zhang, K.M., Ying, Q., Liang, J., Kleeman, M., Wexler, A., Kaduwela, A., 2005a. Particulate matter modeling in central and northern California. In: The Fourth Annual CMAS Models-3 User's Conference, Chapel Hill, NC, 26–28 September 2005.

Zhang, Y., Vijayaraghavan, K., Seigneur, C., 2005b. Evaluation of three probing techniques in a three-dimensional air quality model. Journal of Geophysical Research 110, D02305.